Uncertainty-Based Referral for Selective Chest X-Ray Classification

Nada A. Attia, Walid Al-Atabany, and Sahar Selim

Center for Informatics Science (CIS), School of Information Technology and Computer Science, Nile University, 26th of July Corridor, Sheikh Zayed City, Giza, 12588, Egypt
NElhady@nu.edu.eg, sselim@nu.edu.eg

Abstract. Automated chest X-ray classifiers are increasingly used for triage and decision support, but fixed-threshold inference can be unsafe when models produce overconfident errors or uncertain predictions. We study uncertainty-aware selective prediction for multi-label frontal chest radiograph classification using frozen MedSigLIP image features and lightweight prediction heads. We compare four uncertainty signals: MC-dropout variance, MC-dropout entropy, ensemble variance, and ensemble predictive entropy, alongside random-referral and confidence-based baselines. Study-level uncertainty is computed from per-label uncertainty scores, and the most uncertain studies are referred to expert review while the remaining cases are handled automatically. On a five-label CheXpert frontal-view task, ensemble predictive entropy achieves the highest macro-F1 at the 20% referral operating point, improving macro-F1 from 0.9265 to 0.9387 while reducing false negatives from 972 to 661, an approximately 32.0% reduction, and false positives from 1572 to 1122 on the automated subset. These findings suggest that uncertainty estimation can provide an actionable referral signal for human-in-the-loop chest X-ray classification.

Keywords: uncertainty estimation · selective prediction · chest X-ray · multi-label classification · human-in-the-loop AI

1 Introduction

Chest radiography is one of the most widely used imaging examinations, and automated chest X-ray interpretation has become an important benchmark for clinical decision support [1]. However, deployment in clinical workflows requires more than strong average performance. In triage and reporting settings, automated systems must reduce workload while avoiding unsafe decisions, especially when false negatives may delay care and radiographic labels can be ambiguous or incomplete. Standard multi-label classifiers usually convert probabilities into binary decisions using a fixed threshold, forcing predictions even for borderline or unreliable studies. In addition, modern deep networks can be miscalibrated and may produce overconfident errors, making raw probabilities insufficient for

deciding when predictions should be trusted [2]. These challenges motivate selective prediction systems that decide not only what label to predict, but also when a study should be deferred to expert review.

Uncertainty estimation provides a natural mechanism for selective deployment. Monte Carlo (MC) dropout estimates uncertainty using multiple stochastic forward passes at inference time [3], while deep ensembles estimate uncertainty through disagreement across independently trained models [5]. These repeated predictions can be summarized using disagreement-based measures such as predictive variance, or predictive uncertainty measures such as entropy. In medical imaging, such signals can support reject-option or learning-to-defer settings, where high-uncertainty cases are routed to clinicians and low-uncertainty cases remain eligible for automation [8, 6, 7]. However, the practical role of uncertainty estimation for multi-label chest X-ray deployment remains insufficiently characterized. In particular, it is unclear which lightweight uncertainty signal is most useful for referral, how referral affects false negatives and false positives on the automated subset, and whether observed gains are robust under statistical resampling. Recent studies have examined related questions in uncertainty benchmarking and deferral. Baur et al. [12] benchmark uncertainty disentanglement in multi-label chest X-ray classification, focusing on the behavior of different uncertainty components in CXR prediction. Wundram and Baumgartner [13] compare uncertainty-based deferral with learned deferral in medical AI. In contrast, this study focuses on lightweight study-level referral for frontal CheXpert classification using frozen MedSigLIP features, explicit random and confidence baselines, FN/FP error trade-offs, Brier score, and bootstrap robustness at clinically interpretable referral rates.

In this work, we present a validation-driven selective deployment study for multi-label frontal chest X-ray classification, where uncertainty is evaluated by its ability to support safer automation rather than by visualization alone. Our contributions are: (i) a systematic comparison of MC-dropout variance, MC-dropout entropy, ensemble variance, and ensemble predictive entropy under a common frozen-feature pipeline; (ii) a conservative study-level referral policy based on maximum per-label uncertainty; and (iii) an operating-point evaluation at 20% referral that measures macro-F1, false negatives, false positives, Brier score, and bootstrap resampling-based statistical robustness.

2 Method

2.1 Dataset and Labels

Experiments are conducted on the CheXpert chest radiography benchmark [1]. We use frontal-view radiographs and construct a patient-level train/validation/test split to avoid patient overlap across splits. Because only frontal radiographs are considered, each evaluation instance contains a single frontal image from one study. The resulting dataset contains 152,594 training images, 19,442 validation images, and 18,991 test images. The task is multi-label classification of five findings: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion.

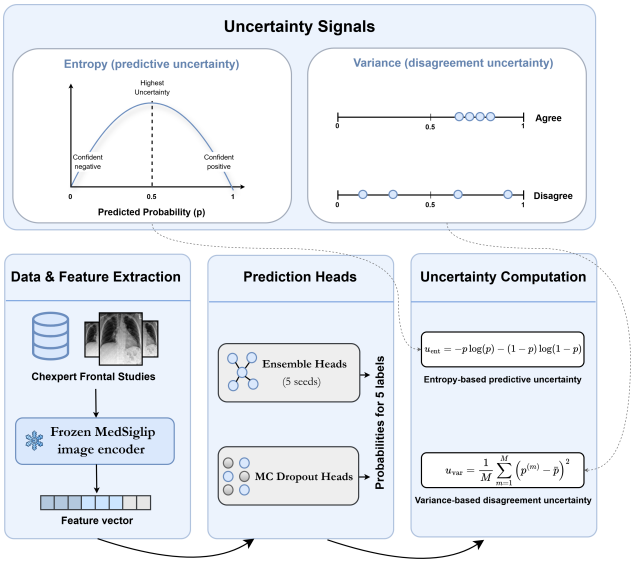

Fig. 1. Overview of the validation-driven selective prediction framework. Frontal CheXpert radiographs are encoded with a frozen MedSigLIP image encoder, and lightweight heads produce multi-label probabilities for five findings. MC-dropout and ensemble predictions are used to compute entropy- and variance-based uncertainty scores. Per-label uncertainties are aggregated into a study-level score, and the most uncertain studies are referred to expert review. The primary operating point is 20% referral, corresponding to 80% automatic coverage.

Positive and negative labels are used as binary targets, while uncertain and missing labels are excluded using a per-label validity mask. All losses and metrics are computed only over valid, unmasked labels.

2.2 Backbone and Feature Extraction

We use the MedSigLIP vision encoder (`google/medsiglip-448`) as a frozen feature extractor [10]. For each radiograph x, the visual encoder produces a pooled representation $\mathbf{h}(x)$, which is passed to a lightweight multi-label prediction head for the five target findings:

$$\mathbf{p}(x) = \sigma(f_\theta(\mathbf{h}(x))). \tag{1}$$

Only the visual encoder is used; the language branch is not used. The backbone remains frozen in all experiments, so uncertainty estimation and referral are evaluated as a lightweight deployment layer.

2.3 Prediction Heads and Uncertainty Signals

We evaluate two lightweight mechanisms for obtaining repeated predictions from the frozen MedSigLIP features: MC dropout and deep ensembles. For MC dropout,

dropout is applied within the prediction head and kept active at inference time, producing T stochastic probability estimates:

$$\mathbf{p}^{(t)}(x) = \sigma\left(f_{\theta,t}(\mathbf{h}(x))\right), \qquad t = 1, \ldots, T. \tag{2}$$

The MC predictive mean is computed as

$$\bar{\mathbf{p}}(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}^{(t)}(x). \tag{3}$$

For the ensemble setting, we train M independently initialized prediction heads and obtain member predictions:

$$\mathbf{p}^{(m)}(x) = \sigma\left(f_{\theta_m}(\mathbf{h}(x))\right), \qquad m = 1, \ldots, M. \tag{4}$$

The ensemble predictive mean is computed as

$$\bar{\mathbf{p}}(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{p}^{(m)}(x). \tag{5}$$

For both MC dropout and ensembles, we compute two types of per-label uncertainty. First, predictive entropy measures uncertainty in the final mean probability:

$$u_j^{\text{ent}}(x) = -\bar{p}_j(x)\log\left(\bar{p}_j(x)\right) - \left(1-\bar{p}_j(x)\right)\log\left(1-\bar{p}_j(x)\right). \tag{6}$$

Second, predictive variance measures disagreement across repeated predictions:

$$u_j^{\text{var}}(x) = \frac{1}{K}\sum_{k=1}^{K}\left(p_j^{(k)}(x) - \bar{p}_j(x)\right)^2. \tag{7}$$

Here, $K = T$ for MC dropout and $K = M$ for ensembles. Entropy is used as a predictive uncertainty score, while variance is used as a disagreement-based uncertainty score. This gives four uncertainty signals: MC-dropout variance, MC-dropout entropy, ensemble variance, and ensemble predictive entropy.

Implementation details. For reproducibility, all prediction heads were trained on frozen MedSigLIP features using a lightweight MLP head with hidden dimension 512 and dropout probability 0.2. The ensemble used five independently initialized heads, and MC-dropout used 20 stochastic forward passes at inference time. Heads were trained for up to 50 epochs with early stopping patience 8, AdamW optimization, learning rate 1×10^{-3}, weight decay 1×10^{-4}, training batch size 2048, and binary cross-entropy loss over valid, unmasked labels. Inference and referral evaluation used batch size 4096. Experiments were run on an NVIDIA RTX 8000 GPU, and all referral thresholds were selected on the validation split only.

2.4 Study-level Referral Policy

Because chest X-ray classification is multi-label, uncertainty is first computed per label and then aggregated into a single study-level score. For each study, we use the maximum uncertainty across the five findings:

$$u_{\text{study}}(x) = \max_{j \in \{1,\ldots,C\}} u_j(x).$$

This maximum aggregation implements a conservative and clinically motivated referral rule: a study is referred whenever any clinically relevant finding has high uncertainty. This is appropriate for multi-label chest X-ray classification because uncertainty in even one target finding may affect clinical interpretation, so the study should be escalated rather than automatically accepted.

For a referral rate r, the referral threshold is defined on the validation set as

$$\tau_r = \text{Quantile}_{1-r}\left(\{u_{\text{study}}(x_i)\}_{i=1}^{N_{\text{val}}}\right). \tag{8}$$

A study is referred to expert review if

$$u_{\text{study}}(x) \geq \tau_r. \tag{9}$$

Otherwise, it is assigned to the automated subset, denoted as AUTO. For the confidence baseline, uncertainty is computed directly from the deterministic predicted probability as $u_j^{\text{conf}}(x) = \min(\bar{p}_j(x), 1 - \bar{p}_j(x))$, so predictions closer to 0.5 are treated as more uncertain. The same maximum study-level aggregation and referral-rate thresholding are then applied. For the random-referral baseline, 20% of test studies are selected uniformly at random for referral, independent of model probability or uncertainty. This is repeated over 1000 random draws, and the reported results are the mean performance over the remaining automated subsets. We evaluate referral rates of 5%, 10%, 15%, 20%, and 30%. The primary operating point is set to 20% referral, corresponding to 80% automatic coverage. This operating point was chosen as a practical trade-off between safety and workload.

2.5 Automated Prediction and Evaluation

Automated predictions are produced only for studies that remain in the AUTO subset after referral. Unless otherwise specified, binary decisions are obtained using a fixed classification threshold of 0.5:

$$\hat{y}_j(x) = \mathbf{1}\left[\bar{p}_j(x) \geq 0.5\right], \qquad x \in \text{AUTO}. \tag{10}$$

We do not use post-referral threshold tuning in the primary setting, so that improvements can be attributed to the referral policy and the choice of uncertainty signal rather than to additional decision-threshold optimization.

Performance is evaluated using macro-F1, false positives, false negatives, and Brier score, with metrics computed using the per-label validity mask described in Sec. 2.1. To assess robustness, we use bootstrap resampling over the test set and compare uncertainty methods under the same 20% referral operating point.

Statistical analysis. Statistical robustness was assessed using 1000 bootstrap resampling iterations over the test set at the 20% referral operating point. In each bootstrap iteration, macro-F1 was recomputed for each uncertainty method on the automated subset. Pairwise statistical comparisons were performed using paired Wilcoxon signed-rank tests comparing each method against ensemble predictive entropy. To account for multiple comparisons, p-values were adjusted using the Holm-Bonferroni procedure, with statistical significance defined as adjusted $p < 0.05$. The resulting significance markers are reported in the bootstrap comparison figure.

3 Results

3.1 Selective prediction performance at 20% referral

Table 1 reports test-set performance at the primary 20% referral operating point, corresponding to 80% automatic coverage. The no-referral baseline evaluates all studies with a fixed threshold of 0.5, whereas selective prediction defers the most uncertain 20% of studies to expert review and evaluates the remaining automated subset using the same threshold. A random-referral baseline with 1000 draws tests whether gains exceed simply deferring 20% of studies.

Random referral remains close to no referral (macro-F1 0.9260), while the confidence baseline is strong (macro-F1 0.9373, lowest FP). MC-dropout variance gives the weakest improvement, while entropy-based signals perform better. Ensemble variance achieves the lowest false-negative count, but with more false positives. Ensemble predictive entropy achieves the highest macro-F1 and reduces false negatives from 972 to 661, corresponding to an approximately 32.0% reduction.

All selective results are reported at the predefined 20% referral operating point, corresponding to 80% automatic coverage, as defined in Sec. 2.4. Across referral rates of 5-30%, ensemble predictive entropy macro-F1 increased from 0.9286 to 0.9455; Fig. 2 shows the trend across all uncertainty signals.

Table 1. Test-set comparison at the primary 20% referral operating point. The no-referral baseline evaluates all studies, while selective methods evaluate the 80% automated subset after deferring studies. Random referral results are reported as the mean over 1000 random referral draws. Macro-F1 values include 95% confidence intervals.

Method	Coverage	Macro-F1 [95% CI]	FN	FP	Brier
No referral	100%	0.9265 [0.9228, 0.9299]	972	1572	0.0729
Random referral	80%	0.9260 [0.9241, 0.9278]	800.4	1247.3	0.0732
Confidence baseline	80%	0.9373 [0.9337, 0.9412]	688	**1111**	0.0655
MC dropout variance	80%	0.9320 [0.9279, 0.9351]	736	1227	0.0711
MC dropout entropy	80%	0.9374 [0.9338, 0.9408]	694	1108	0.0656
Ensemble variance	80%	0.9372 [0.9336, 0.9410]	**654**	1248	0.0685
Ensemble predictive entropy	80%	**0.9387 [0.9350, 0.9423]**	661	1122	**0.0655**

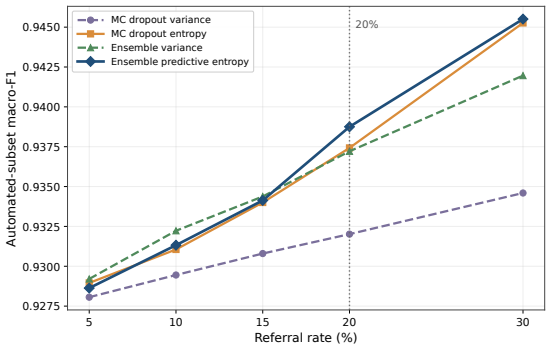

Fig. 2. Automated-subset macro-F1 across referral rates. The vertical dotted line marks the primary 20% referral operating point.

3.2 Error trade-off across uncertainty signals

Figure 3A summarizes false negatives and false positives at the 20% referral operating point. The no-referral baseline has the highest absolute error counts because it evaluates all studies. Random referral also reduces absolute FN/FP counts by evaluating fewer studies; however, Table 1 shows that confidence- and uncertainty-based referral achieve stronger macro-F1 and Brier scores than random referral. This indicates that the selection strategy contributes to the improvement beyond reduced coverage alone.

The uncertainty signals show different error profiles. MC-dropout variance gives the weakest improvement, while MC-dropout entropy provides stronger false-positive reduction. Ensemble variance achieves the lowest false-negative count (654), but with more false positives (1248). Ensemble predictive entropy achieves the highest macro-F1 and a strong false-negative profile, reducing false negatives to 661 and false positives to 1122.

3.3 Bootstrap robustness and statistical comparison

Bootstrap resampling was used to assess robustness by recomputing macro-F1 for no referral and the four uncertainty-based methods over the test set; random and confidence baselines are reported separately in Table 1. As shown in Figure 3B, the no-referral baseline has the lowest macro-F1 distribution, while all four uncertainty-based referral methods improve automated-subset performance. Entropy-based methods show stronger distributions than MC-dropout variance, consistent with Table 1.

Ensemble predictive entropy achieves the highest and most stable macro-F1 distribution, remaining consistently above the no-referral baseline. Together with the error-count analysis, these results show that the improvement is not limited to macro-F1 alone: the selected referral signal also reduces both false negatives and false positives on the automated subset.

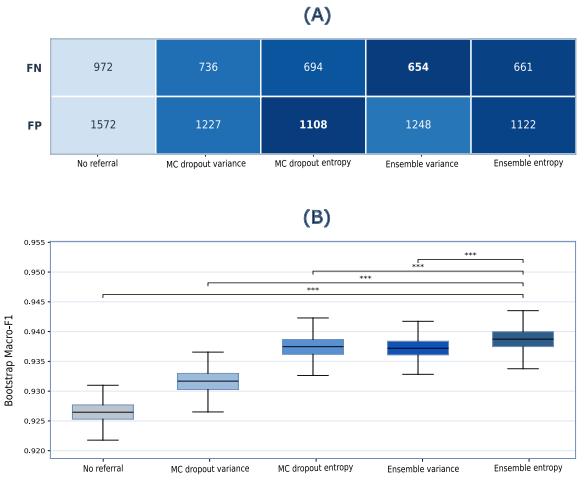

Fig. 3. Error trade-off and statistical robustness of uncertainty-based referral at 20% referral. (A) False negatives and false positives; darker cells indicate lower error within each row. (B) Bootstrap macro-F1 distributions for no referral and the four uncertainty-based methods; random and confidence baselines are reported in Table 1. Significance markers compare each uncertainty method against ensemble predictive entropy using Holm-corrected p-values: $* \; p < 0.05$, $** \; p < 0.01$, $*** \; p < 0.001$.

4 Discussion

This work frames uncertainty estimation for chest X-ray classification as a selective deployment problem rather than a post-hoc visualization task. We compared multiple lightweight uncertainty signals under the same frozen-feature pipeline and evaluated whether they improve automated-subset reliability by deciding which studies should be handled automatically and which should be deferred to expert review.

The main finding is that the choice of referral signal affects the safety-coverage trade-off. MC-dropout variance gave the weakest improvement, while entropy-based signals produced stronger gains. Ensemble variance achieved the lowest false-negative count, but with more false positives. The confidence baseline was also highly competitive, achieving a macro-F1 close to ensemble predictive entropy, the same Brier score, and slightly fewer false positives. Ensemble predictive entropy achieved the highest macro-F1 and a lower false-negative count than the confidence baseline, making it preferable to the confidence baseline when false-negative reduction is prioritized.

The results also emphasize the need for deployment-relevant evaluation. Conventional discrimination metrics are useful but insufficient when the goal is to improve reliability on the automated subset. We therefore focus on macro-F1, false negatives, false positives, Brier score, and bootstrap resampling-based statistical comparisons. The bootstrap analysis supports the robustness of the ob-

served gains and shows that the selected uncertainty signal is not favored only by a single point estimate.

The selected 20% referral rate should be interpreted as a predefined deployment operating point rather than a universal optimum. Different clinical settings may choose different coverage targets depending on radiologist capacity, workload constraints, and acceptable safety requirements.

The framework is lightweight because it operates on frozen image features and prediction heads, but this also limits uncertainty estimation to the classifier layer rather than the full image encoder. The experiments are limited to CheXpert frontal-view studies with masked uncertain and missing labels, so external validation under dataset shift is needed. Finally, real clinical deployment would require prospective evaluation with radiologists to measure workflow impact and decision quality.

Overall, uncertainty-aware selective prediction provides a practical mechanism for safer chest X-ray automation: lower-uncertainty studies can remain eligible for automated prediction, while uncertain cases are escalated to expert review.

5 Conclusion

This paper studied uncertainty-aware selective prediction for multi-label chest X-ray classification using frozen MedSigLIP features and lightweight prediction heads. We compared MC-dropout variance, MC-dropout entropy, ensemble variance, and ensemble predictive entropy as referral signals for identifying studies that should be deferred to expert review.

On a five-label CheXpert frontal-view task with masked uncertain and missing labels, selective referral consistently improved automated-subset performance. At 20% referral, corresponding to 80% automatic coverage, ensemble predictive entropy achieved the highest macro-F1 among the evaluated methods while substantially reducing both false negatives and false positives relative to no referral. These results suggest that uncertainty-based referral can support safer human-in-the-loop chest X-ray deployment by reserving automated predictions for lower-uncertainty studies and escalating uncertain cases to expert review.

References

1. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In: Proc. AAAI. pp. 590-597 (2019)
2. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks. In: Proc. ICML. pp. 1321-1330 (2017)
3. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Proc. ICML (2016)

4. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
5. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
6. Geifman, Y., El-Yaniv, R.: SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In: Proc. ICML. Proc. of Machine Learning Research (PMLR), vol. 97, pp. 2151-2159 (2019)
7. Mozannar, H., Sontag, D.: Consistent Estimators for Learning to Defer to an Expert. In: Proc. ICML. Proc. of Machine Learning Research (PMLR) (2020)
8. Chow, C.K.: On Optimum Recognition Error and Reject Tradeoff. IEEE Transactions on Information Theory **16**(1), 41-46 (1970)
9. Borah, J., Singh, H.K.: DCAT: Dual-view Cross-Attention Transformer for Multi-label Chest X-ray Classification with Uncertainty Estimation. arXiv preprint arXiv:2503.11851 (2025)
10. Google Health AI Developer Foundations: MedSigLIP model card (2025), `https://developers.google.com/health-ai-developer-foundations/medsiglip/model-card`.
11. Kahl, K.-C., Lüth, C.T., Zenk, M., Maier-Hein, K., Jaeger, P.F.: ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation. In: Proc. International Conference on Learning Representations (ICLR) (2024)
12. Baur, S., Samek, W., Ma, J.: Benchmarking Uncertainty and its Disentanglement in Multi-label Chest X-Ray Classification. arXiv preprint arXiv:2508.04457 (2025)
13. Wundram, A.M., Baumgartner, C.F.: Is Uncertainty Quantification a Viable Alternative to Learned Deferral? arXiv preprint arXiv:2508.02319 (2025)